III. On certain Points connected with scale Order in the Case of the Correlation of two characters which for some arrangement give a Linear Regression Line.

By KARL PEARSON, F.R.S.

In a recent memoir on contingency*, I have considered the problem of what alterations can be made in scale order without sensibly modifying the value of the correlation. The problem as I there state it is as follows: *To find under what other condition than normal correlation small changes in the order of grouping will not affect the value of the correlation* (p. 19). The wording requires some explanation. If for any arrangement of the scales of the two variables there be normal correlation, then my memoir shows that the method of contingency gives the value of the correlation, even if the order of the scales be any whatever, in fact if the normal correlation order be absolutely unknown. Of course, if we proceed in any such case by the usual product method of determining the correlation we shall reach absolutely different results when the scale order of grouping is largely changed. My object in stating the above problem was to determine, if possible, whether any and if so what changes in the scale orders would not sensibly modify the correlation, when we still endeavoured to determine it, not by contingency, but by the method of products. The conclusion I came to was as follows—that with any distribution *with linear regression* "small changes (i.e. such that the sum of their squares may be neglected as compared with the square of mean or standard deviation) may be made in the order of grouping without affecting the correlation coefficient" (p. 35). I think this conclusion is quite sound, and deserves further consideration. Although in the statement of the proposition I have used the word "small changes" in scale order (p. 19) and in the summary of my memoir (p. 35) stated what is to be understood by small, in this case, I think, as Mr G. U. Yule points out to me, that the wording on p. 20 is too unguarded, if the reader has not been sufficiently impressed with the wording on p. 19, or reached the summary on p. 35. It will not be without value possibly to give the actual algebraical result on which the statement on p. 35 is based, for it has some importance for the general philosophical idea of correlation.

Let x and y represent the two variable characters and let $u\delta x$ be the frequency of the character between x and $x+\delta x$; $v\delta y$ that of the character between y and $y+\delta y$; u and v being functions of x and y respectively and the distribution of the frequencies being of any nature. Now suppose the array $v_s\delta y_s$ of frequency between y_s and $y_s+\delta y_s$ to be bodily interchanged in position with the array $v_{s'}\delta y_{s'}$ between $y_{s'}$ and $y_{s'}+\delta y_{s'}$. Let N be the total frequency, and suppose the mean $\bar{y}$ to become $\bar{y}+\delta\bar{y}$, the standard deviation σ_y of the y character to become $\sigma_y+\delta\sigma_y$. Then we have:

$$N(\bar{y}+\delta\bar{y}) = S(yv\delta y) - v_{s'}\delta y_{s'}(y_{s'}-y_s) - v_s\delta y_s(y_s-y_{s'})$$

or
$$\delta\bar{y} = (y_s-y_{s'})\frac{(v_{s'}\delta y_{s'}-v_s\delta y_s)}{N} \quad \text{.....(i),}$$

$$N(\sigma_y+\delta\sigma_y)^2 = S(y^2v\delta y) - v_{s'}\delta y_{s'}(y_{s'}^2-y_s^2) - v_s\delta y_s(y_s^2-y_{s'}^2) - N(\bar{y}+\delta\bar{y})^2$$
$$= N\sigma_y^2 + (v_{s'}\delta y_{s'}-v_s\delta y_s)(y_s^2-y_{s'}^2) - 2\bar{y}(y_s-y_{s'})(v_{s'}\delta y_{s'}-v_s\delta y_s),$$
$$N(\delta\sigma_y)^2 + 2N\sigma_y\delta\sigma_y = (v_{s'}\delta y_{s'}-v_s\delta y_s)(y_s-y_{s'})(y_s-\bar{y}+y_{s'}-\bar{y}).$$

* "Mathematical Contributions to the Theory of Evolution, III. On the Theory of Contingency and its Relation to Association and Normal Correlation." *Drapers' Research Memoirs* (Dulau and Co. London).

Hence we see that $\delta\sigma_y$ is small, if the frequencies of interchanged subgroups are small as compared with N and accordingly:

$$\delta\sigma_y/\sigma_y = \frac{v_{s'}\delta y_{s'} - v_s\delta y_s}{N}\,\frac{(y_s - y_{s'})}{\sigma_y}\,\frac{y_s - \bar{y} + y_{s'} - \bar{y}}{2\sigma_y} \quad \text{.....(ii).}$$

We now turn to the change in the product-moment.

$$P + \delta P = S(wxy\delta x\delta y) - v_{s'}\delta y_{s'}\bar{x}_{s'}(y_{s'} - y_s) - v_s\delta y_s\bar{x}_s(y_s - y_{s'}) - N\bar{x}(\bar{y} + \delta\bar{y}),$$

where $w\delta x\delta y$ is the total frequency of individuals, with characters between x and $x+\delta x$ and y and $y+\delta y$ and $\bar{x}_s$ and $\bar{x}_{s'}$ are the means of the arrays corresponding to y_s and $y_{s'}$. But $P = S(wxy\,\delta x\delta y) - N\bar{x}\bar{y}$, hence:

$$\delta P = (y_s - y_{s'})\left((\bar{x}_{s'} - \bar{x})v_{s'}\delta y_{s'} - (\bar{x}_s - \bar{x})v_s\delta y_s\right).$$

Thus
$$\delta P/P = \frac{y_s - y_{s'}}{\sigma_y}\left(\frac{\bar{x}_{s'} - \bar{x}}{r\sigma_x}\,\frac{v_{s'}\delta y_{s'}}{N} - \frac{(\bar{x}_s - \bar{x})}{r\sigma_x}\,\frac{v_s\delta y_s}{N}\right) \quad \text{.....(iii).}$$

Now if r be the correlation before and $r+\delta r$ after a change is made, we have, since $r = P/(N\sigma_x\sigma_y)$,

$$\frac{\delta r}{r} = \frac{\delta P}{P} - \frac{\delta\sigma_x}{\sigma_x} - \frac{\delta\sigma_y}{\sigma_y} \quad \text{.....(iv).}$$

Now we have supposed at present no change to be made in the x's; thus we may treat $\delta\sigma_x$ as zero, and using (ii) and (iii) we have, rearranging:

$$\frac{\delta r}{r} = \frac{y_s - y_{s'}}{r\sigma_y\sigma_x}\left[\frac{v_{s'}\delta y_{s'}}{N}\left\{\bar{x}_{s'} - \bar{x} - \frac{r\sigma_x}{\sigma_y}(y_{s'} - \bar{y})\right\} - \frac{v_s\delta y_s}{N}\left\{\bar{x}_s - \bar{x} - \frac{r\sigma_x}{\sigma_y}(y_s - \bar{y})\right\}\right]$$
$$- \frac{(y_s - y_{s'})^2}{2\sigma_y^2}\,\frac{v_{s'}\delta y_{s'} + v_s\delta y_s}{N} \quad \text{.....(v).}$$

Now suppose the regression to be originally linear, then we have $\bar{x}_s - \bar{x} = \frac{r\sigma_x}{\sigma_y}(y_s - \bar{y})$ not only for s and s' but for all values of s whatever. In other words the whole series of terms in square brackets vanishes and summing for all pairs of interchanges:

$$\frac{\delta r}{r} = -\frac{S(y_s - y_{s'})^2(v_{s'}\delta y_{s'} + v_s\delta y_s)}{2N\sigma_y^2} \quad \text{.....(vi).}$$

If we make similar interchanges of x_p and $x_{p'}$ we can show that*:

$$\frac{\delta r}{r} = -\frac{S(y_s - y_{s'})^2(v_{s'}\delta y_{s'} + v_s\delta y_s)}{2N\sigma_y^2} - \frac{S'(x_p - x_{p'})^2(u_{p'}\delta x_{p'} + u_p\delta x_p)}{2N\sigma_x^2}$$
$$+ \frac{S'''(y_s - y_{s'})(x_p - x_{p'})(w_1\delta x_p\delta y_s - w_2\delta x_{p'}\delta y_s - w_3\delta x_p\delta y_{s'} + w_4\delta x_{p'}\delta y_{s'})}{Nr\sigma_x\sigma_y} \quad \text{...(vi) bis.}$$

Here S denotes a summation or integration for all possible interchanges of the y arrays, i.e. say, columns of the correlation table; and S' denotes a like summation for all possible interchanges of the x-arrays, say the rows of the table. S''' is a summation involving the frequency at all points where interchanged rows and columns cross. Of course this result assumes that the units of grouping of both characters are so "fine" that the squares of the ratios of the array frequencies to the total frequency are negligible.

We may now draw some interesting conclusions from (vi). Suppose the material to be such that the correlation is linear under some arrangement. Then for slight interchanges the squares and products of the interchanges are negligible and δr will be zero. Thus, r being positive, we

* The reader will find a verification of this formula arising from writing (i) the correlation table with its columns inverted, then $\delta r/r = -2$, and (ii) again in addition with its rows written backwards, in this case $\delta r/r = 0$. In (i) the first term only remains and its numerator $= 4N\sigma_y^2$. In the second case the numerators of the three terms are respectively $4N\sigma_y^2$, $4N\sigma_x^2$ and $4Nr\sigma_x\sigma_y$.

see from (vi) that r is an absolute maximum. Clearly $\delta r/r$ is always negative even for interchanges between arrays at considerable distances. Or, we conclude that if there be one arrangement of the material for which the regression line is linear, then any interchanges, however extensive, will reduce the value of the correlation as calculated by the product moment method. This conception of the linear regression line as giving the arrangement with the maximum degree of correlation appears of considerable philosophical interest. It amounts practically to much the same thing as saying that if we have a fine classification, we shall get the maximum of correlation by arranging the arrays so that the means of the arrays fall as closely as possible on a line.

Further, if the mean square of the interchanges, i.e. the expression

$$\frac{S(y_s - y_{s'})^2 (v_{s'}\delta y_{s'} + v_s \delta y_s)}{2N},$$

be small as compared with the standard deviation squared, i.e. σ_y^2, then the change δr will not be sensible. In other words *small* changes in the scale ordering, not confined to adjacent or even to two arrays, will not sensibly modify the correlation as found by the product moment method.

Lastly, considering the proof of (vi) we see that no portion of the investigation is dependent on the whole of the one y-array being interchanged with the whole of another. We may consider $v_s\delta y_s$ and $v_{s'}\delta y_{s'}$ as only portions of the total array—to be taken, however, proportionately from all its constituents. Now let $V_s\delta y_s$ and $V_{s'}\delta y_{s'}$ denote the whole of the frequency of the two arrays, and write the first array $V_s\delta y_s + \frac{1}{2}m - \frac{1}{2}m$ and the second array $V_{s'}\delta y_{s'} - \frac{1}{2}m + \frac{1}{2}m$. Now transfer the $-\frac{1}{2}m$ of the first array to the position of the second and the $+\frac{1}{2}m$ of the second to the position of the first, i.e. take $v_s\delta y_s = -\frac{1}{2}m$ and $v_{s'}\delta y_{s'} = +\frac{1}{2}m$; it follows that $v_s\delta y_s + v_{s'}\delta y_{s'} = 0$ and the two arrays are

$$V_s\delta y_s + m \text{ and } V_{s'}\delta y_{s'} - m,$$

i.e. exactly the values they would have had if a portion of the second array drawn at random from all its sub-groups had been inscribed in the same sub-groups of the first array. But in this case we see since $v_s\delta y_s + v_{s'}\delta y_{s'} = 0$, that (vi) will give us absolutely $\delta r = 0$, or there will be no change in the correlation. This result seems of considerable value. Suppose the regression linear, and one character, x say, easily measured or known; then if a number m of individuals which ought to fall into a given class of y, be shifted by oversight or error of judgment into a second erroneous class of y, this will not sensibly affect the correlation, if N being the total frequency, the square of the ratio m/N is negligible, as compared with its first power. Thus suppose in correlating age with hair tint, the first character being accurately known, an observer were to place his series of contributory observations of hair tint in the wrong group, say in one of the brown reds instead of pure browns, this would not sensibly modify the resulting correlation. The fact that the error would not produce a modification is not in the first place due to the possible smallness of the misplaced group. The product moment is changed and the standard deviation is also modified, but the modification of the correlation depends on such manner on the changes of these two, that they act in opposite senses and cancel the modification, provided the original regression was strictly linear.

While not desiring to encourage carelessness in observing or tabling or in the formation of scale orders without due consideration, still the results of this note seem to indicate that in many cases absolute unanimity of judgment in classifying or great stress on small details of scale grouping are not needful in order to reach sensibly identical values of the correlation. This view coincides with my actual and not unique experience, when having been in grave doubt as to where 30 or 40 individuals were to be placed, I put them first in one category and then in a second, only to find out that the correlation worked out with the group first in one and then in the other category was sensibly identical. The theorems developed in this note seem to explain this stability—when we use not contingency but product moment methods, and suppose the regression ultimately linear.